A meta-model for representing bioinformatics workflow to improve reproducibility

Mouna EL GARB[1], Emmanuel COQUERY[1], Fabien DUCHATEAU[1] and Nicolas LUMINEAU[1]

Université Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622 Villeurbanne, France

Corresponding author: mouna.el-garb@univ-lyon1.fr

Abstract *This poster presents a meta-model for improving reproducibility in scientific workflows.*

Background

Computational workflows in Snakemake, Nextflow, or Galaxy have become essential tools in the bioinformatics field for managing and analyzing the increasing amount of data produced in this domain [1,2]. In addition to better portability, scalability and re-entrancy, the execution steps and intermediate results of a workflow can be shared and reused in new workflows. However, achieving this requires a structured and fine-grained description of the workflow using metadata and provenance in a generic, machine-readable, and workflow-engine-independent format [3]. Several specifications have been proposed to describe workflows: the P-Plan ontology for provenance and plans [4], which extends the W3C PROV-O ontology [5], RO-Crate which is able to capture the provenance of workflows [6], and SWCF, a provenance model to capture the behaviour of control dependencies [7]. However, these representations remain limited, as no model fully addresses all aspects of workflows. Key elements, such as sub-workflows, detailed information about scripts used at each step (e.g., arguments and libraries), and the specific commands executed by tools are not explicitly defined. Additionally, the naming conventions used to represent workflow information can be ambiguous. Furthermore, control flow is often represented in a complex manner in the proposed models.

Results

To tackle these issues, we propose BioFlow-Model, a meta-model that integrates all relevant concepts introduced in main ontologies, along with new concepts to cover all the aspects of workflows. In addition to common concepts (description of workflow steps, their interdependencies, the data they use and generate, and the data flow within the workflow), our model includes a detailed representation of control flow and provenance trace elements (useful for workflows driven by the control flow using control patterns such as parallel split, as proposed in NextFlow). Where possible, our model extends existing classes and properties from existing models, thus reusing concepts that were already effective. Our BioFlow-Model, available on 10.5281/zenodo.14945693, consists of 19 classes (7 new ones and 12 reused or mapped to existing classes), and 34 properties (4 new ones, and 30 reused or mapped to existing properties).

Conclusion

The poster session will provide an opportunity to discuss our meta-model with the bioinformatics community, including users or contributors of existing models. As this model serves as the foundation of an upcoming workflow query language, we will also illustrate typical use cases for retrieving relevant (parts of) workflows.

References

[1] Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature methods. 2021;18(10):1161-8.

[2] Leipzig J. A review of bioinformatic pipeline frameworks. Briefings in Bioinformatics. 2016 03;18(3):530-6. Available from: https://doi.org/10.1093/bib/bbw020.

[3] Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Generation Computer Systems. 2017;75:284-98. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X17300316.

[4] Garijo Verdejo D, Gil Y. Augmenting prov with plans in p-plan: scientific processes as linked data. In: Proceedings of the Second International Workshop on Linked Science. CEUR Workshop Proceedings; 2012. p. 1-4. Available from: http://purl.org/net/p-plan.

[5] Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, et al. Prov-o: The prov ontology. W3C recommendation. 2013;30. Available from: https://www.w3.org/TR/prov-o/.

[6] Leo S, Crusoe MR, Rodríguez-Navas L, Sirvent R, Kanitz A, De Geest P, et al. Recording provenance of workflow runs with RO-Crate. PLoS one. 2024;19(9):e0309210. Available from: https://www.researchobject.org/ro-crate/specification/1.1/index.html.

[7] Anila Sahar Butt PF. A provenance model for control-flow driven scientific workflows. Data Knowledge Engineering. 01/2021;131-132.

Acknowledgements

This work was funded by a government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-22-PESN-0007 (ShareFair project).